公開

（事後評価・研究者用）

研　究　報　告　書

「パラフレーズ現象の解明のための言語資源構築とパラフレーズアラインメント技術の確立」

研究期間：　2018 年 4 月～2020 年 3 月
研究者番号：　50201
研 究 者：　荒瀬　由紀

1.　研究のねらい

本研究では、文法的なフレーズを単位とするパラフレーズを対象とし、パラフレーズ研究において必須となる言語資源の構築と、それを応用した高精度なテキストのベクトル化技術およびフレーズアラインメント技術の開発に取り組む。

言語は表現の自由度が高く、同じ事象であっても多様な語彙・構文構造をもって記述できる。このような多様性は人工知能による言語理解における高い障壁である。言語理解が求められるアプリケーションとして、自動質問応答や Google Home のようなパーソナルアシスタントがある。これらアプリケーションでは、ユーザの問いかけを理解し処理するため、知識ベースを用いることが多い。本質的には知識ベースにある知識を使って回答可能であっても、言語的な表現の違いにより知識ベースとのマッチングが上手くいかず、誤った回答を行う、回答そのものが不可能となるといった問題が起こる。このような同一の事象を異なる言語表現で表す現象をパラフレーズと呼ぶ。パラフレーズの理解および知的な処理は、言語理解を要するアプリケーションにおける基盤であり、最重要課題の一つである。

これまでパラフレーズ認識は盛んに研究されてきたが、パラフレーズか否かという二値の分類性能向上が主要課題であった。これは、パラフレーズ認識に失敗するものにおいてどのような言語現象が要因となっているのか分析を可能とするデータセットが存在しなかったことが要因である。またテキストのベクトル表現は言語理解の基盤となる技術であるが、実アプリケーションに応用可能なレベルの言語理解を実現するには、さらなる技術革新が望まれている。

そこで本研究では（1）パラフレーズ文ペアにおけるフレーズアラインメントのアノテーションにより、パラフレーズ抽出研究に必須となる言語資源を構築し、それを活用して（2）高精度なテキストベクトル化技術の開発、および（3）柔軟なフレーズアラインメント技術の開発、に取り組む。

2.　研究成果

（1）概要

文法的なフレーズを単位とするパラフレーズを対象とし、パラフレーズ研究において必須となる言語資源の構築と、それを応用した高精度なテキストベクトル化技術、フレーズアラインメント技術の開発に取り組む。これらは人間の話す言語を人工知能が理解する上での基盤となるものであり、自動質問応答やパーソナルアシスタントなど、幅広い応用を持つ。本研究では（1）パラフレーズ文ペアにおけるフレーズアラインメントのアノテーションにより、パラフレーズ

認識研究に必須となる言語資源を構築し、それを活用して（2）高精度なテキストベクトル化技術および（3）柔軟なフレーズアラインメント技術を開発する。

まず（1）では、英語母語話者である言語学者に依頼し、パラフレーズ文対に構文構造のアノテーションを実施する。これにより、文法なフレーズの抽出が可能となる。そして得られたフレーズ対について、英語母語話者およびバイリンガル話者 3 名によりフレーズアラインメントのアノテーションを行う。合計 1,916 文対についてアノテーションを完了しており、フレーズアラインメントのアノテーションデータとしては世界最大のデータを構築した。

（2）では、深層学習を用いた高精度なテキストベクトル化技術を開発した。パラフレーズ認識を含め様々な文対モデリングタスクに応用し、性能を体系的に評価したところ、既存手法を大きく上回る性能を達成している。本成果は自然言語処理分野の最重要国際会議の一つである Empirical Methods in Natural Language Processing（EMNLP 2019）にて採択されている（主要研究成果 5－(1)-2）。

（3）では、シンプルかつ柔軟なフレーズアラインメント技術を開発した。これまで我々が開発してきたフレーズアラインメント技術は高精度なアラインメントを行える一方、Head-driven phrase structure grammar（HPSG）による構文解析を必要としており、適用範囲が限定的であった。そこでテキストのベクトル化手法を用い、HPSG 構文木に縛られない柔軟なフレーズアラインメント技術を開発した。

（2）詳細

研究テーマ（1）パラフレーズ文ペアにおけるフレーズアラインメントのアノテーション

パラフレーズ 1,916 ペアについて、まず文法的なフレーズ抽出を可能とするため、英語を母語とする言語学者により構文構造のアノテーションを行った。その後、構文情報を用いて得られるフレーズ 15.2 万件について、英語母語話者およびバイリンガル話者 3 名により 2 つのフレーズ対が同一の意味を表現するかどうかラベル付けを行った。その結果、約 25.2 万件のアラインメントを獲得し、内ユニークなアラインメントは約 10.5 万件であった。アノテータ 3 名の内、2 名を正解、1 名を評価対象とした場合の、アラインメントの再現率は 93.3%、適合率は 90.2%、F 値は 91.7% であった。意味の同一性判定は本質的に曖昧性が存在するが、今回得られたアノテーション結果は充分高い一致率をもつといえる。構築したアノテーションデータの統計量を表 1 に示す。本データセットは文法的フレーズアラインメントをラベル付けしたものとして世界最大であり、付随する構文構造を用いてパラフレーズ認識手法の詳細なエラー分析が可能である。構築したデータセットは Linguistic Data Consortium より公開予定である。

表 1 フレーズアラインメントデータセットの統計量

パラフレーズ文対数	1,916
語彙サイズ	9,540
フレーズ総数（単語を除く）	152,480（75,283）
アラインメント数（ユニークなアラインメント数）	251,972（105,154）
2 名以上のアノテータで一致したアラインメント数	80,572

研究テーマ（2） 高精度なテキストベクトル化技術の開発

テキストのベクトル表現は言語理解の基盤となる技術であり、現在活発に研究が行われている分野である。既存研究では深層学習によりあらゆるタスクに適用可能な汎用的なベクトルの生成を目指しており、性能を高めるためにモデルの大規模化を行ってきた。しかし数十億のパラメータを持つような大規模なモデルは、時間的・空間的計算量の要件からアプリケーションへの応用が難しいというジレンマを抱えている。

本研究では言語理解の基盤となる、パラフレーズ認識や含意関係認識等の文対モデリングにフォーカスし、それらに特化したテキストのベクトル化を小規模なモデルで実現する手法を開発した。文対モデリングにおいて構文的な情報やフレーズの意味的関係は重要な役割を果たすと考えられる。しかし既存のテキストベクトル化手法では文における構文構造やフレーズ間の意味的関係は一切考慮されていなかった。そこで、汎用的な文ベクトル生成を行う BERT (Devlin et al. 2019) にフレーズアラインメントの訓練を加える。具体的には図 1 に示す通り、提案手法では BERT が生成する単語ベクトルをプーリングしたものをフレーズベクトルとして、フレーズアラインメントが可能かどうか分類する学習を行う。これにより、フレーズに関わる構文的知識を BERT に付加し、かつ意味的類似性をベクトルに反映することを可能とする。推論時は BERT レイヤのみを用いてテキストのベクトル化を行う。つまり、提案手法では BERT のパラメータを増やすことなくテキストのベクトル化性能を向上する。

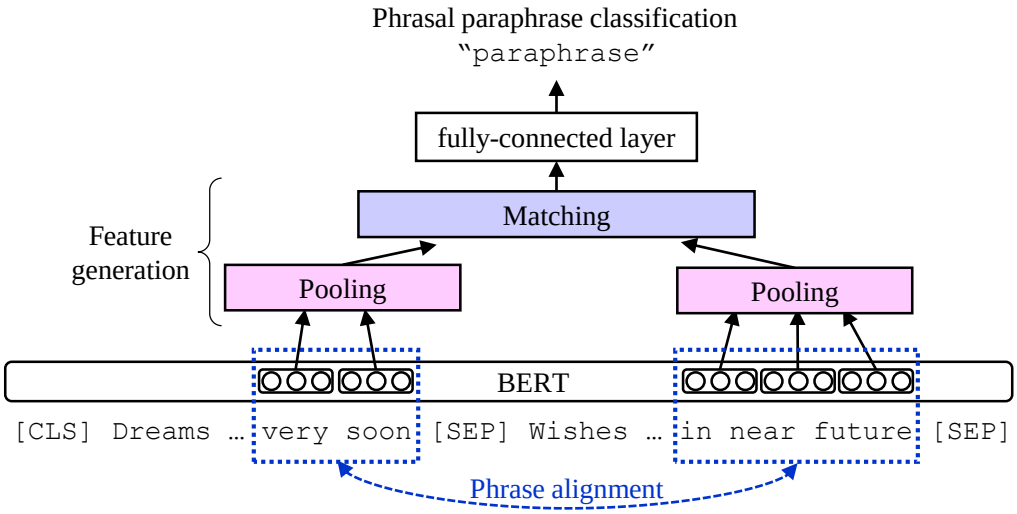

図 1 フレーズアラインメントを応用したテキストベクトル化

表 2 にパラフレーズ認識、意味的類似度推定、含意関係認識タスクにおける提案手法および BERT の性能を示す。各手法はタスク固有のラベル付き訓練データによって Fine-tuning を行い、テストセットで性能を評価する。BERT は標準モデルである base モデル、その 3.1 倍のパラメータを持つ large モデルを用い、それぞれを訓練した提案手法（Ours-base および Ours-large）と比較する。表 2 が示す通り、提案手法ではそれぞれ対応する BERT を上回る性能を達成している。さらに意味的類似度推定タスクの STS-B、含意関係認識タスクの RTE、パラフレーズ認識タスクの MRPC では、Ours-base が 3.1 倍のパラメータを持つ BERT-large と同等もしくはより高い性能を示している。以上

より提案手法がモデルサイズを維持しながらも、様々な文対モデリングタスクの性能を向上するベクトル化を行うことが示された。

表 2 文対モデリングタスクにおける性能（最高性能を太字で、対応する BERT より高い性能のものをアンダーラインで示す）

	パラフレーズ認識		類似度推定	含意関係認識			
Model	MRPC	QQP	STS-B	MNLI-m	MNLI-mm	QNLI	RTE
BERT-base	88.3	71.2	84.7	84.3	83.0	89.1	59.8
Ours-base	<u>88.6</u>	<u>71.5</u>	**87.7**	<u>84.7</u>	<u>83.6</u>	<u>91.1</u>	<u>67.0</u>
BERT-large	88.6	72.1	86.0	86.2	85.5	**92.7**	65.5
Ours-large	**89.9**	**72.5**	<u>87.1</u>	**86.5**	**85.6**	92.2	**68.2**

また 図 2 に、STS-B タスクの訓練データサイズを変動させた際の開発セットにおける性能を示す。グラフより、提案手法は訓練データが少ない際に特に大きな性能向上を示すことが分かる。他のタスクにおいても同様の傾向が確認され、提案手法は構築コストが高いタスク固有の訓練（ラベル付き）データを削減できることが明らかとなった。

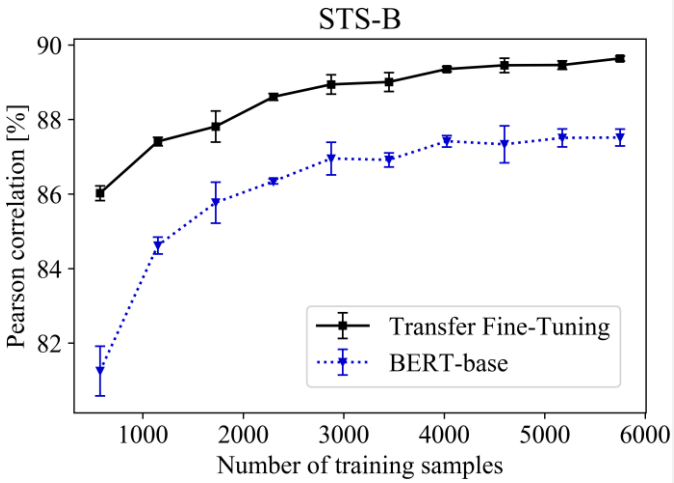

図 2 訓練データサイズの影響（黒線が提案手法）

研究テーマ（3） 柔軟なフレーズアラインメント技術の開発

我々がこれまで開発してきたフレーズアラインメント手法（Arase and Tsujii 2017）は、高精度なアラインメントが可能な一方、HPSG 構文解析器に強く依存していた。構文解析器の精度はドメインの影響を受けやすく、専門性の高い文や口語体の文の解析精度は大きく低下することが知られている。そこで本研究では構文木の構造（文法）に縛られず、あらゆるフレーズのアラインメントを可能とする手法を開発した。テキストベクトル化モデルによりフレーズをベクトル化し、ラティスを枝刈りしながらフレーズアラインメントを行うことでフレーズの構造に対し柔軟なアライメントを高速に行う（主要研究成果 5－(1)-1, 5－(3)-1）。

参考文献

Y. Arase and J. Tsujii. 2017. Monolingual Phrase Alignment on Parse Forests, in Proc. of EMNLP, pp. 5396-5407.

Jacob Devlin et al. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. in Proc. of NAACL-HLT, pp. 4171-4186.

3. 今後の展開

パラフレーズ認識は人工知能による言語理解における基盤技術である。今後は開発したテキストベクトル化技術、フレーズアラインメント技術を対話システムに応用する予定である。同一の意味を表現する入力を知的に考慮することで、ユーザとシステムの対話が破綻しないよう対話システムの頑健性を高める。

またパラフレーズ研究の重要テーマとして、パラフレーズの自動生成がある。パラフレーズ生成は、意味を保持しつつテキストの難易度を柔軟に変更するテキスト平易化、文体を変更するスタイル変換など、特に言語教育に貢献する技術である。既存のパラフレーズ生成技術には、意味の同一性を保持できず過剰な言い換えを行う、あるいは意味の同一性を保持しようとするあまりほとんど言い換えを行わないなど、課題が多く残されている。今後、これまで取り組んできたフレーズアラインメント技術を応用して品質の高いパラフレーズ生成技術の開発に取り組む予定である。

4. 自己評価

研究目的の達成状況

本研究では文法的なフレーズを単位とするパラフレーズを対象とし、パラフレーズ研究において必須となる言語資源の構築と、それを応用した高精度なテキストのベクトル化技術およびフレーズアラインメント技術の開発に取り組んだ。成果として世界最大となるフレーズアラインメントデータセットを構築し、またそれを活用したテキストのベクトル化技術、柔軟なフレーズアラインメント技術の開発を行った。技術面では既存の state-of-the-art を上回る性能を達成し、その成果は EMNLP 等重要国際会議に採択されている。またフレーズアラインメントの改善手法についても論文投稿を準備中である。以上より、研究目的は充分達成され、また当初予定を超えて研究を進めることができたと考えている。

研究の進め方（研究実施体制及び研究費執行状況）

研究費を活用してフレーズアラインメントアノテーションを行い、学術的にも貴重なデータセットを構築できた。データセットについては学術利用のため、近日中に公開予定である。

研究成果の科学技術及び学術・産業・社会・文化への波及効果

パラフレーズ認識は言語理解における基盤技術であり、自動質問応答や対話システム等、人工知能と人間の言葉によるインタラクションを実現する上で必須となる技術である。それらに貢献する本研究は学術的にも社会的にも重要な成果である。今後、本研究で開発した技術を実際に対話システムや自動質問応答システムに適用し、これらアプリケーションの性能向上に貢献したい。

研究課題の独創性・挑戦性

文法的なフレーズを単位とするパラフレーズに関する研究はほとんど前例がなく、また本研究は言語資源構築から新たな技術の開発まで体系的に行う挑戦性の高いプロジェクトであった。アドバイザや他の ACT-I 生からの助言、活発な議論がなければ本プロジェクトを完遂することはできなかった。心より感謝申し上げたい。

5. 主な研究成果リスト

(1)論文(原著論文)発表

1. M. Yoshinaka, T. Kajiwara, and Y. Arase. SAPPHIRE: Simple Aligner for Phrasal Paraphrase with Hierarchical Representation. in Proc. of International Conference on Language Resources and Evaluation (LREC 2020), (May 2020 to appear).
2. Y. Arase and J. Tsujii: Transfer Fine-Tuning: A BERT Case Study, in Proc. of Conference on Empirical Methods in Natural Language Processing (EMNLP2019), pp. 5396-5407 (Nov. 2019).
3. Y. Arase and J. Tsujii: SPADE: Evaluation Dataset for Monolingual Phrase Alignment, in Proc. of Language Resources and Evaluation Conference (LREC 2018), (May 2018).

(2)特許出願

(3)その他の成果(主要な学会発表、受賞、著作物、プレスリリース等)

1. 吉仲真人, 梶原智之, 荒瀬由紀: 単語分散表現に基づく単一言語内フレーズアライメント手法. 言語処理学会第 26 回年次大会, pp.581-584, (2020 年 3 月).
2. 荒瀬由紀. 人と AI を言葉でつなぐ～自然言語処理による言語理解～, 「AI と人がつくる未来社会」シンポジウム (日本学術会議第三部, 近畿地区会議, 国立大学法人大阪大学) (2019 年 8 月). (招待講演)
3. SPADE. LDC Catalogue No.: LDC2018T09 (https://catalog.ldc.upenn.edu/LDC2018T09).